哥伦比亚大学口述历史中心项目聚类分析与主题解构*

王阮 邓君 孙绍丹 宋先智 宋雪雁
吉林大学管理学院 长春 130022

摘 要：[目的/意义]以全新视角审视挖掘哥伦比亚大学口述历史中心(CCOH)项目特征，深入解构项目主题，为丰富我国口述历史项目主题提供参考基准。[方法/过程]选取哥伦比亚大学口述历史中心 9 586 个口述历史项目为研究对象，利用 Python 抓取数据，通过 Word2Vec 和 K-means 对筛选后的 3 589 个项目(过滤无简介项目 5 997 个)聚类分析并进行项目主题解构。[结果/结论]哥伦比亚大学口述历史中心项目包含 4 个主题层面：①以人物为主题——兼容并包，辐射各行各业群体；②以机构为主题——博采众长，发挥自身职能优势；③以地区为主题——统筹规划，联动地区体系建设；④以特定时期、具体运动/事件为主题——创新范式，观照家国时代主题。

关键词：CCOH 项目主题 Word2Vec K-means
分类号：G270
DOI：10.13266/j.issn.0252-3116.2020.17.004

1 引言

作为口述历史研究的肇始之源，哥伦比亚大学口述历史研究室[1]的建立奠定了美国口述历史长足发展的基础。此后，口述历史在加拿大、英国、法国、新加坡、日本等国家得到迅速发展，涌现出一大批口述历史研究者和专业研究团体。2011 年 6 月 1 日，该研究室正式更名为哥伦比亚大学口述历史中心(Columbia Center for Oral History，CCOH)，经过 70 年的发展，CCOH 研究范围不断深入扩展，几乎渗透到各个领域，并在逐渐探索中形成自己的研究特色，成为世界口述历史研究的标杆机构，与加州大学伯克利分校口述历史办公室、加州大学洛杉矶分校口述历史研究中心并称为美国三大口述历史中心，是从事口述历史实践和教学的世界领先中心之一[2]。至此，美国第一批以高校为基础的口述历史发展格局已基本形成。

中国口述历史实践最早可追溯至 1980 年国家科学院自然科学史研究所发起的口述科技史项目。相比之下，在起点上落后美国 30 多年。近年来，伴随着对非物质文化遗产的关注升温，口述历史受到持续重视，我国相继成立了温州大学口述历史研究所、崔永元口述历史研究中心等机构，旨在更好地传承和保护口述历史资源，然而我国口述历史项目实践进程较为缓慢且资源存储丰富度与资源描述细粒度略显匮乏，项目主题单一。基于此，本文择取口述历史实践的先驱——哥伦比亚大学口述历史中心 3 589 个口述历史项目(共 9 586 个项目，过滤无简介项目 5 997 个)为研究对象[3]，利用 Python 爬取项目简介，采用 Word2Vec (Word to Vector，也被称为"Word Embed-ding")和 K-means 对项目聚类分析并进行主题解构，通过对人物、机构、地区和特定时期事件等主题回顾和梳理，以全新视角审视并深入挖掘项目主题特征，希冀收获启示一二。

2 哥伦比亚大学口述历史中心的缘起与沿革

哥伦比亚大学口述历史研究室由历史学家 A. Nevins 于 1948 年创办[4]。成立之初，该研究室以收集日记和信件等文件资料为主。随着信息技术不断发展，口述访谈成为主要手段，初期的采访以政坛要物为主要对象。此后，研究范围逐渐拓展至其他领域，诸如慈善组织、媒体、商务、艺术、社会群体等[5]。自 2011 年

* 本文系国家社会科学基金项目"数字人文视角下历史档案资源知识聚合与知识发现研究"(项目编号：19BTQ102)研究成果之一。
作者简介：王阮(ORCID：0000-0002-5445-4943)，博士研究生；邓君(ORCID：0000-0002-3291-7193)，教授，博士生导师，通讯作者，E-mail：dengjun9722@163.com；孙绍丹(ORCID：0000-0003-2927-0971)，博士研究生；宋先智(ORCID：0000-0002-7499-2431)，博士研究生；宋雪雁(ORCID：0000-0003-1955-6685)，教授，博士生导师。
收稿日期：2020-03-23 修回日期：2020-05-30 本文起止页码：37-48 本文责任编辑：王传清

该研究室更名为哥伦比亚大学口述历史中心后,2013 年 11 月,为进一步区分该中心的研究和档案功能,分别设立 CCOHR(Columbia Center for Oral History Research)[6] 和 CCOHA(Columbia Center for Oral History Archives)[7]。目前,CCOH 已收录超过 8 000 个文本和音频采访[8],其资源丰富度和主题多元化拓宽了国内外口述历史研究视域。

通过文献阅读可知,国内学者对 CCOH 研究涉猎较为广泛。从发展脉络看,学者陈水湘[9]、王玉龙[10]、刘佳敏[11]对 CCOH 发展概况做了简要梳理;从职能分析看,立足宏观视角,张洪涛对 CCOH 岗位设置与人员职能、资源建设与加工组织、资源存档与使用、教育培训与跨学科合作、资金来源与调查研究 5 个方面进行了详细调查[5]。聚焦细粒度微观视角,在资源采集层面,CCOH 以"自主采访为主,合作采集 + 捐赠为辅"模式进行项目建设[12],并且 CCOH 的工作人员不直接参与口述史料的采集,而是通过招募有经验的采访者策划和完成访谈的全过程;在组织建设层面,CCOH 成立了由哥伦比亚大学相关院系、研究所和图书馆领导以及专家学者组成的顾问委员会,负责口述历史教育培训和研究工作[13];在管理与操作层面,CCOH 尤为严谨规范,其在口述历史项目建设过程中严格遵守美国口述历史协会(Oral History Association)、伦理审查委员会(Institutional Review Board)以及知识共享组织(Creative Commons)有关访谈操作和法律伦理规范等方面的详细规定[14]。在数字化存储与检索层面,由于 CCOH 设立在哥伦比亚大学巴特勒图书馆,因而巴特勒图书馆作为美国口述历史资源数字化中心,在处理口述音频和抄本及相关资料的数字化方面为口述历史项目资源库建设奠定了基础[15]。同时,CCOH 与哥伦比亚大学图书馆数字保存团队联合设计的口述历史在线门户(Oral History Portal)可以实现对口述历史馆藏、口述历史项目以及个别口述历史访谈的关键词检索[16]。在传播推广层面,CCOH 不仅提供哥伦比亚大学图书馆珍贵手稿的开放利用,而且还通过专门的网上商店出售口述历史相关缩微制品[17];或通过舞台剧、展览、动画等多种形式展现项目成果,与公众积极良性互动。在资金筹措方面,CCOH 大部分项目得到了美国国家科学基金会的资助与奖励[18]。

虽然也有国外学者[19]记录 CCOH 发展历程,但国外学者多围绕 CCOH 某一特定项目,或对项目完整流程予以揭示[20],或揭露项目所反映的社会现实,如对劳动、种族、阶级和性别的关注[21-24],对特别事件的保存记载[25]。

综上,学者们多采用网络调查、文献分析等传统方法对 CCOH 展开调研,研究方法略显匮乏,鲜少注入创新性研究方法。从研究内容看,现有文献为 CCOH 研究奠定了良好的理论基础,同时对推进 CCOH 实践具有一定启发意义。国内学者对 CCOH 的研究既侧重理论层面探讨(如机构发展概况、人员配置介绍、职能分析)又注重实践流程探索(如资源采集、管理与操作、数字化存储与检索、传播推广),然而在实践进程中学者们过多聚焦运转流程,疏于对 CCOH 口述历史项目主题的关注,欠缺对项目主题的细粒度解读;而国外学者虽然也对 CCOH 发展予以记录,较为侧重项目主题研讨,但项目样本单一,无法囊括项目总览,辐射项目全景。

因此,笔者希望以全新视角审视 CCOH 项目,引入 Word2Vec 和 K-means 方法,对过滤后得到的 3 589 个项目聚类分析并进行主题解构,试图从人 - 事 - 地 - 机构四维层面深入剖析项目主题特性,挖掘项目内涵,为优化我国口述历史资源配置,丰富项目主题提供可靠依据,对我国制定口述历史项目实践战略具有深远意义。

3 研究方法

3.1 Word2Vec

Word2Vec 是 Google 2013 年开源的一款词向量训练工具,该算法包含 CBOW(continuous bag-of-words model)和 Skip-Gram(continuous skip-gram model)两种模式,前者根据上下文推测特征词,后者通过特征词推测上下文概率。Word2Vec 在主题聚类领域具有极强的适用性和灵活性,能够从大规模未经标注的语料中高效生成词向量,并且针对篇幅较长的文章在关键词提取的准确性方面明显高于其他算法[26]。

3.2 K-means

K-means 起源于 1957 年,由 S. Lloyd 提出。相比于其他算法,该算法适合处理大容量数据集[27],聚类集成更加高效[28],依赖参数取值大小,尤其是对聚类 K 值的选择,决定实验结果好坏,并且支持在没有任何先验知识情况下,对多个属性聚类分析,包含以下 4 个步骤:

第一步:随机选取 k 个样本均值点,记第 i 个均值为 u_i。

第二步:求取各个样本点到各均值点的距离,把距离最短的归为一类;$j = 1、2、\cdots、k$,$i = 1、2、\cdots、n$,d 为第

i 个点到第 j 个均值距离，x_i 为第 i 个数据点：

$$d = \sqrt{\sum_{i=1}^{n}(x_i - u_j)^2} \qquad 公式(1)$$

第三步：对第二步得到的新的 k 类，分别求取均值，得到新的均值点：

$$u_k = \frac{\sum_{i=1}^{nk} x_i}{nk} \qquad 公式(2)$$

第四步：重复第二、第三步骤，直到操作得到的样本均值点不再显著变化为止。

鉴于 Word2Vec 和 K-means 两种算法操作简单、高效，因而不乏有学者将两者结合应用于某领域主题聚类分析。如 H. Tao 等采用 Word2Vec 模型强化话题语义信息表达，设置加权 K-means 算法进行话题识别[29]。沈思等应用 Word2vec 和 K-means 对食品安全事件相关微博和信息数据应急处理，探究了类别知识挖掘的具体流程、方法和相应的类别知识挖掘结果评价等问题[30]。由此可知，两种方法的结合使用被广泛应用于各领域，且实验效果较好。

本研究首先利用 Python 抓取数据源，对文本进行 NLTK 分词预处理，然后将 Word2Vec 和 K-means 算法结合使用进行 CCOH 项目聚类分析，具体来说，Word2Vec 将文本转化为向量表示，随后作为 K-means 集成模型数据输入，获得 K 值，即聚类结果，以此获得项目主题个数，进一步进行主题解构。

4 哥伦比亚大学口述历史中心项目聚类分析

4.1 数据来源及预处理

本研究利用 Python 抓取数据源为 CCOH 收录的 9 586个口述历史项目，笔者对项目进行整理，结果如图 1 所示：

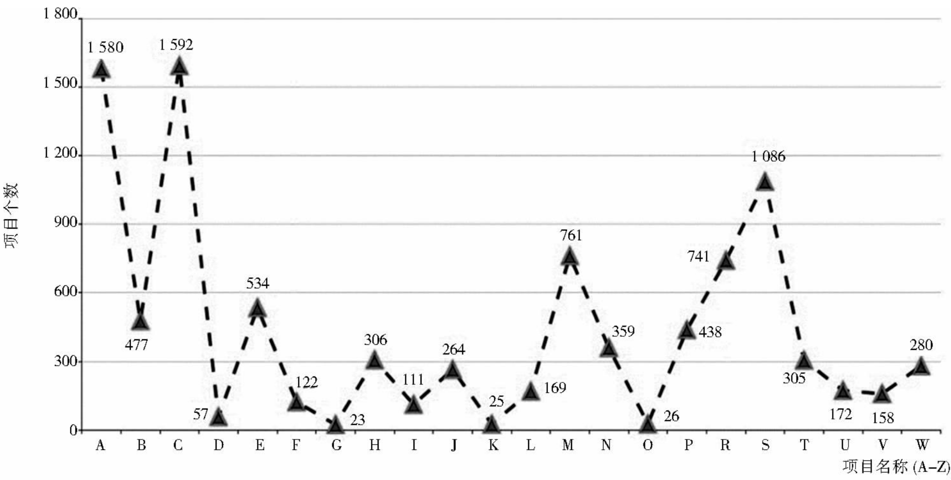


图 1 哥伦比亚大学口述历史中心(CCOH)口述历史项目统计

将爬取数据存储在文本编辑器 Sublime Text，利用的第三方库为 Requests 和 XML，通过 Python 的 Scrapy 爬虫框架采集项目背景简介，保存为. CSV 格式。部分代码如图 2 所示：

```
html = html.xpath('//*[@id="midwrapper"]/table[2]/tbody/tr/td[1]/div[%s]/p[2]/a[1]/@href' %j)
projects.append(html.xpath('//*[@id="midwrapper"]/table/tbody/tr/td[1]/div/table/tbody/tr[1]/td[2]/strong/text()'))
contents.append(html.xpath('//*[@id="midwrapper"]/table/tbody/tr/td[1]/blockquote[1]/text()'))
```

图 2 Python 抓取目标网站数据部分代码展示

4.2 高频词及词云图

利用 Python 对所爬取的数据源做词频统计，对经 NLTK 分词包预处理后的文本数据进行大小写合并、同义词合并后得到高频词表(见表 1)。

由表 1 可知，排名靠前且频次较高的高频词便于识别 CCOH 口述历史项目主题，对研究具有启发性。为更加清晰呈现项目主题词，增强视觉效果，笔者利用 WordArt 生成高频词云图，如图 3 所示。CCOH 口述历史项目以“interview”形式开展，据此推测每个项目均围绕这一基点对口述历史资源进行采集、组织与描述、存储与利用、共享及推广，藉以深化 CCOH 口述历史项目研究的深度与广度。同时，“family”“childhood”“life”“education”“background”等词频围绕访谈背景展开叙述，“president”“Eisenhower”“engineer”“worker”“Black”“White”等词频展现出访谈对象身份变化，“corporation”“administration”“university”“association”

表 1　CCOH 口述历史项目词频统计表(频次 >100)

高频词	频次	高频词	频次	高频词	频次	高频词	频次
education	949	corporation	274	administration	178	marine	123
worker	741	personal	272	officer	177	board	120
university	617	research	266	academy	176	assistant	120
war	531	background	264	role	174	social	116
association	514	company	261	city	171	law	116
school	487	politic	260	Columbia	162	teaching	115
family	452	service	254	Institute	158	reminiscence	115
childhood	420	life	243	art	144	business	113
interview	400	national	241	organization	139	state	110
oral history	384	Black	238	involvement	138	plant	109
experience	367	navy	230	Africa	138	White	108
relation	359	training	228	engineer	137	production	108
American	353	general	209	studies	136	Washington	107
career	344	press	198	operation	130	health	107
Carnegie	323	campaign	195	committee	129	International	106
development	318	Eisenhower	194	staff	129	army	105
continental	304	discuss	191	Bennington	127	policy	105
president	296	department	186	foundation	126	Turkey	104
college	295	student	185	industry	124	study	104

"company"等词频体现了组织机构的多样性,"Africa""Turkey""Washington"等词频折射出地域的广泛性,"war""campaign"等词频凸显了战争背景。故而,高频词的词频变化从不同层面反映了 CCOH 口述历史项目主题特征,蕴含在诸多项目的语义揭示中。



图 3　CCOH 口述历史项目高频词云图

4.3　哥伦比亚大学口述历史中心口述历史项目聚类分析

聚类分析是为了更加清楚地了解 CCOH 项目主题,探寻项目特征。本文利用 NLTK 对预处理后的数据进行分词处理,同时基于 Gensim 包中的 Word2Vec 和 K-means 算法对项目聚类分析,实验工具为 jupyter notebook。

4.3.1　数据清洗

分词停用词处理是文本分析数据处理的关键步骤,对文本分析的准确性至关重要。鉴于中英文书写

王阮，邓君，孙绍丹，等．哥伦比亚大学口述历史中心项目聚类分析与主题解构[J]．图书情报工作，2020，64(17)：37－48．

格式的差异性，在提取特征词前，本研究首先对数据进行分词处理，去除无用符号，采用 NLTK 分词预处理，新增 split 列存入分词停用词处理后的数据以适应文本精确模式，如图 4 所示：

	i	x	split
0	1	Mao Zedong's early career, his own experience...	Mao Zedong career , experiences studying abroa...
1	2	Ma Chaojun discusses Nationalist politics and...	Ma Chaojun discusses Nationalist politics pers...
2	3	addresses Shi's experiences through 1936, wit...	addresses Shi experiences 1936 , attention Nor...
3	4	Chen talks about military politics in Zhejian...	Chen talks military politics Zhejiang , involv...
4	5	Fu Bingchang's career as a legal expert and ...	Fu Bingchang career legal expert diplomat Repu...
5	6	his military education, its development, and ...	military education , development , internal Re...
6	7	He discusses his military education and servi...	He discusses military education service Beiyan...
7	8	Hu gives accounts of battles and politics in ...	Hu accounts battles politics unifying Guangdon...
8	9	Ji shares his knowledge of military affairs i...	Ji shares knowledge military affairs Yunnan Ma...
9	10	his political life in Beijing and Guangxi duri...	political life Beijing Guangxi Republican peri...

图 4　分词后数据

4.3.2　词向量构建

利用 Word2Vec 将文本转化成向量，用文本中每个词向量的平均值表示文字向量，以 x1、x2、x3、x4、x5……表示，转化为 Data Frame 格式(见图 5)。

	i	x	split	x1	x2	x3	x4	x5	x6	x7	...	x41	x42	x43	x44	x45	x46	x47	x48	x49	x50
0	1	Mao Zedong's early career, his own experience...	Mao Zedong career , experiences studying abroa...	0.427888	-0.260461	0.148203	-0.177586	-0.036586	0.437527	-0.688265	...	-0.143623	-0.521304	0.064336	0.147741	0.970124	0.067975	-0.496507	-0.044999	0.295844	-0.337494
1	2	Ma Chaojun discusses Nationalist politics and...	Ma Chaojun discusses Nationalist politics pers...	0.427888	-0.260461	0.148203	-0.177586	-0.036586	0.437527	-0.688265	...	-0.143623	-0.521304	0.064336	0.147741	0.970124	0.067975	-0.496507	-0.044999	0.295844	-0.337494
2	3	addresses Shi's experiences through 1936, wit...	addresses Shi experiences 1936 , attention Nor...	0.318518	-0.067654	0.174612	0.022143	-0.117779	0.244916	-0.619790	...	-0.102864	-0.310197	-0.011595	0.050985	0.686243	0.114476	-0.457714	-0.071257	0.162162	0.026002
3	4	Chen talks about military politics in Zhejian...	Chen talks military politics Zhejiang , involv...	0.388696	-0.099439	0.156541	0.010763	-0.105246	0.251121	-0.553923	...	-0.117399	-0.339487	-0.041285	-0.015584	0.642715	0.049827	-0.412298	-0.095318	0.169759	0.033787
4	5	Fu Bingchang's career as a legal expert and ...	Fu Bingchang career legal expert diplomat Repu...	0.354634	-0.128326	0.124554	-0.089556	-0.042344	0.356695	-0.646556	...	-0.135203	-0.372460	0.060822	0.209242	0.785142	0.117760	-0.464351	0.008202	0.182191	-0.315308
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3584	3762	Origin and evolution of the Museum of Design; ...	Origin evolution Museum Design ; financial sup...	0.573676	-0.030402	0.220742	0.002410	-0.173881	0.312825	-0.740225	...	-0.175263	-0.337333	-0.067415	0.080447	0.652417	0.095901	-0.489389	-0.109089	0.118526	0.138001
3585	3763	Organization of International Exhibit of Moder...	Organization International Exhibit Modern Arch...	0.657738	0.117605	0.240896	0.200255	-0.225989	0.166794	-0.700160	...	-0.120970	-0.111610	-0.145933	-0.171116	0.528848	0.037953	-0.476252	-0.202930	0.088310	0.395131
3586	3764	The "Love" paintings; copyright protection for...	The `` Love " paintings ; copyright protectio...	0.753031	-0.186996	0.351887	-0.031416	-0.240378	0.366479	-0.915751	...	-0.251628	-0.470855	-0.134152	-0.144248	0.778500	0.090989	-0.450541	-0.214223	0.244973	0.386874
3587	3765	Abstract Expressionist movement; "New York, th...	Abstract Expressionist movement ; `` New York ...	0.924223	-0.009594	0.407468	0.302394	-0.107590	0.097023	-1.046387	...	-0.124815	-0.431566	-0.011218	-0.030557	0.421617	0.229255	-0.452092	-0.126426	0.025024	0.190520
3588	3766	Commission for St. Peter's Church, New York, N...	Commission St. Peter Church , New York , N.Y. ...	0.857852	0.036361	0.380573	0.404172	-0.103858	-0.004910	-0.929772	...	-0.021068	-0.353161	-0.006904	-0.061335	0.305717	0.226187	-0.449147	-0.149841	-0.023072	0.167480

图 5　词向量(Data Frame)

4.3.3　K 值选取

有关 k-means 算法分类 k 值随机选取，目前有两种方法(平均轮廓系数和手肘法)可用于评估最优聚类个数。本文采用手肘法，手肘法公式表示如下：

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \qquad \text{公式(3)}$$

SSE 是误差平方和，代表聚类结果的好坏，C_i 是第 i 个簇，p 是 C_i 的样本点，m_i 是 C_i 的质心。随着聚类数 k 的增大，样本划分会更加精细，同时随着每个簇的聚合程度提高，误差平方和 SSE 会逐渐变小。即当 k 小于真实聚类数时(由于 k 的增大会大幅增加每个簇的聚合程度)，SSE 的下降幅度会很大，而当 k 到达真实聚类数时，再增加 k 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 k 值的亟需增大而趋于平缓，也就是说 SSE 和 k 的关系图是一

个手肘的形状，而这个肘部对应的 k 值就是数据的真实聚类数。

本样本聚类评估，k 值与 SSE 之间的变化关系如图 6 所示。显然，由肘部图可知，该数据集最佳聚类数为 4，故 CCOH 项目主题可归纳为 4 个方面。

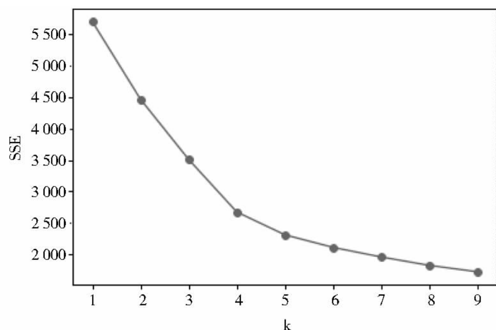


图 6 SSE 肘部图

5 哥伦比亚大学口述历史中心项目主题解构

由聚类结果可知，CCOH 项目主题可以聚为 4 类，此部分将深度解读 CCOH 项目的研究主题，揭示隐含主题特征。

5.1 以人物为主题——兼容并包，辐射各行各业群体

以人物为主题的 CCOH 项目以某些特征或具有社会影响力的个人/群体为中心，着重对人物的鲜活刻画，通过访谈了解人物的职业、性别、年龄、事迹、成果等方面，发现其在各自领域中的贡献轨迹以及该领域未来的发展方向，辐射各行各业。笔者按照人物身份分类，结果如下：

5.1.1 政治类

此类人物多为上层人士，诸如总统、参议员等在内的权威、政治人物。艾森豪威尔项目（Eisenhower Administration Project[31]）记载了艾森豪威尔总统任期（1953 – 1961 年）事迹，访谈对象除艾森豪威尔及家人外，还包括相关白宫人员、内阁成员、政治顾问、国会议员以及大使等。同时，2019 年 7 月 1 日开始，CCOH 正式承接前美国总统、哥伦比亚大学 1983 届校友巴拉克·奥巴马总统口述历史项目（Obama Presidency Oral History Project[32]），该项目是继艾森豪威尔项目后承接的第二个美国总统口述历史项目，将全面记录奥巴马任期所做决策、举措，同时也将揭示米歇尔·奥巴马作为前美国总统第一夫人的相关工作及其产生的影响，目前该项目正在持续推进。对于自身建设之外的口述史，CCOH 自 1958 年开始推进中国名人口述回忆录收藏项目（Academia Sinica Oral History Collection[33]），择取中国各领域重要人物如军事家、哲学家、文学家（如刘景山[34]、贺国光[35]、梅乔林[36]等）。同时，著名历史学家唐德刚撰写整理的李宗仁回忆录也属于中国口述史项目的一部分。

5.1.2 科技类

科技类人物主要以科学家、建筑师、工程师、飞行员等群体为代表。美国工业发展以技术先进、门类齐全、资源丰富、生产实力雄厚、劳动产出率高而著称于世。航空工业发展在一定程度上得益于技术革新突飞猛进，如航空项目（Aviation Project[37]）访谈回忆可追溯至两次世界大战期间航空业的迅速发展，通过访谈工程师、飞行员，以手稿、档案等形式记录滑翔机、轻型飞机、喷气式飞机、超音速飞机的技术变革和发展历程。同航空工业发展相似，矿业发展也源于新技术发明。采矿工程师项目（Mining Engineers Project[38]）深入阿拉斯加和南非地区的矿工生活，充满浓郁生活气息，通过口述访谈识别矿业劳动强度和人物性格特征等，描绘出与众不同的劳动者形象张力和精神感召力。建筑业作为美国经济发展至关重要的发动引擎，同样具有举足轻重的地位。哥伦比亚大学建筑学院启动建筑项目（Architecture Project[39]）助力土木工程发展，选取 4 位伟大建筑创始人（Charles-Edouard Le Corbusier、Walter Gropius、Ludwig Mies van der Rohe、Frank Lloyd Wright）为访谈对象，诠释建筑夯基场景，探寻建筑与城市规划结合点。

随着物质生活不断满足和丰盈，人们的精神世界也随之不断变化。在世界范围内，被公认为所有颁奖领域内最高奖项的诺贝尔奖诞生，从此越来越多的科学家进入公众视野。诺贝尔奖项目（Nobel Laureates on Scientific Research Project[40]）以科学家群体的工作日志和以及家庭生活为主线；托马斯·阿尔瓦·爱迪生项目（Thomas Alva Edison[41]）刻画了爱迪生个人品质与科研历程，形成了 1 048 页的口述访谈稿，成为后人追忆爱迪生的珍贵史料。

5.1.3 民间类

民间类人物以手工艺人最为突出。美国独立之初，几乎无人提及手工艺美术珍品。伴随独立百年（1876 年）纪念日，以口述访谈方式纪念艺术的研究方法非常流行。美国手工艺人项目（American Craftspeople Project[42]）以 5 个特定领域（陶瓷、纤维、玻璃、金属和木材）的手工艺者为访谈对象，"画廊""家具制作""珠宝""陶艺""木雕""编织"等词汇诠释了美国手工

艺发展(特别是过去 30 年间),并试图定义手工艺在美国社会所处地位,揭示随着时间的推移这种角色发生了哪些变化,并研究这些变化产生的时间和原因,有助于人们了解美国丰富多彩的文化背景。

5.1.4 演艺类

演艺类人物涉及演员、导演、摄影师。影视产业是生产与销售以相对独立的物态形式呈现文化产品的行业。电影是一种表演艺术、视觉艺术及听觉艺术,利用胶卷、录像带或数位媒体将影像和声音捕捉,辅以后期编辑而成。1971 年,好莱坞电影产业项目(Hollywood Film Industry Project[43])对好莱坞演员、导演、摄影师和技术人员进行一系列采访,内容涉及制片剪辑、发行出版、电影发展等。从无声到有声,美国逐步诞生了百老汇、好莱坞、迪士尼、时代华纳等大批电影产业巨头,带动了相关产业迅速崛起与发展,逐渐形成了具有鲜明特色的美国文化。

5.1.5 社会类

进入 21 世纪,以人物为主题的项目仍然作为研究主线。社会类人物聚焦企业家、创始人、社会活动家及慈善家等群体。2005 年大西洋慈善口述历史项目(Atlantic Philanthropies Oral History Project[44])第一阶段包含对 139 人的采访,囊括大西洋慈善基金会的领导、Feeney 的商业伙伴、机构合作伙伴和受赠人,访谈重点是记录慈善家 Feeney 的生活。第二阶段,该项目选取 85 人为访谈对象,涉略社会领袖、商界翘楚,通过他们的回忆,获取鲜为人知的慈善事业发展背后故事。

5.1.6 边缘类

不仅如此,本研究还应注意到,以人物为主题的项目不仅仅是上层阶级和各行各业人物的发声物,更是普通民众发展延续的历史见证。CCOH 适时调整视角,将关注焦点转向非裔美国人、墨西哥移民,以及黑人、妇女、劳工等边缘化阶层和社会底层民众,在研究重心的交替中,CCOH 逐渐意识到维护与保存口述历史资料,记录不同阶层发展历史的重要性。

黑人记者口述历史收藏项目(Black Journalists Oral History Collection[45])对 19 世纪末至 70 年代初的非裔美国人有关民权事件、种族隔离做出相关报道。妇女历史和人口问题项目(Women's History and Population Issues Project[46])访谈对象包含与母婴保健服务、婚姻咨询和计划生育相关的妇女。委内瑞拉劳工项目(Venezuelan Labor Leaders Oral History Collection[47])秉持扎根于生活、为劳工而歌的理念,以独特视角揭示出底层劳动者形象。

由此可见,以人物为主题的项目以兼容并包的态势映射出各行各业群体的多元化,辐射领域的广泛化,在丰富、饱满中不乏厚重和灵动,注重人物个体与群体间的和谐统一,折射出 CCOH 不囿于本土人物,同时也在着力挖掘丰富的中国史料,并为兴趣研究者提供了强有力的中国史料补充,且访谈受众群体实现了由上层政治人物向社会无话语权的边缘阶层人物过渡的转变,纵向挖掘历史脉络,传承民族记忆,可以说,这种扁平化拉伸为纵向化的独特视角切换有助于 CCOH 更好地把握社会动向,稳定平衡发展。

5.2 以机构为主题——博采众长,发挥自身职能优势

以机构为主题的 CCOH 项目包含高校、公司、协会、科研院所、基金会等组织,通常以机构沿革为主线,记录机构变迁与发展历史。

5.2.1 高校/公司项目

高校项目立足学术研究与文化传承,如纽约大学项目(New York University Project[48])以著书出版方式记录了纽约大学发展史。公司项目往往立足本司职能分析。作为美国最大的百货商店,联邦百货商店项目(Federated Department Stores Project[49])深入调查消费者群体和消费者购买意愿并归档记录组织发展;世界银行项目(World Bank Project[50])以分析银行管理流程为主,包括政策制定、项目评估、贷款管理、股权与风险投资。

5.2.2 协会/科研院所项目

协会项目聚焦变迁轨迹的场景描绘。美国物理教师协会计划(American Association of Physics Teachers Project[51])记载了 1930 年以来该协会的沿革历史。随着协会不断创新发展,基于发展需求,科研院所也积极启动项目研究。作为美国第一个致力于俄罗斯和苏联跨学科研究的学术中心,哈里曼研究所项目(Harriman Institute Oral History Collection[52])以俄罗斯学术研究和政策状况为背景,记录了该研究所 70 年来的变化。

值得关注的是,宗教活动作为西方生活不可缺少的重要组成部分不容忽视,因此,反映宗教历史的项目建设也是研究院所的重要选题,如克莱蒙特研究生院启动传教士计划(China Missionaries Project[53])协助教派基督徒回忆了该院传教士活动。

5.2.3 基金会项目

通常来说,因背后有强大的财团支撑,所以基金会项目往往实力最强,领域最广,规模最大。纽约卡内基公司于 1911 年成立,是美国历史最悠久,最具影响力的基金会之一。卡内基基金会项目(Carnegie Corpora-

tion Project)分为三部分:第一部分记载基金会 1966 – 1970 年公司发展历程[54];第二部分记载 1996 – 2000 年公司资助贫困黑人的经济研究以及民主经济发展工作,并概述了 25 年来基金会组织发展结构,以及美国慈善事业向民主过渡时期的影响[55];第三部分研究了公司慈善事业的计划和举措,更广泛地讨论了对突发事件(如卡特里娜飓风)的慈善响应以及在经济衰退时期的捐赠管理[56]。

不难看出,高校、公司、协会、科研院所、基金会等组织机构都开辟了独特的项目建设风格。从自身职能视角出发,或立足学术研究与文化传承,或聚焦职能分析,或描绘变迁轨迹,或记录沿革历史,博采众长,围绕自身发展需求完成专属领域的项目建设。正是得益于完备的组织体系以及完整的组织队伍,才促使 CCOH 研究逐步走向欣欣向荣的局面,知识传播度更广,受众面更宽。

5.3 以地区为主题——统筹规划,联动地区体系建设

CCOH 以地区为主题的项目通常限定某一特定地理区域范围来采集口述历史,围绕地区政治、经济、文化、安全、民生、教育等领域展开,纪实地区发展史。

纽约政治研究项目(New York Political Studies (A) Project[57])围绕纽约政治发展问题与成就,探讨了城市发展问题(New York Political Studies (B) Project[58]),同时聚焦 1949 年纽约大选,记录并涵盖了所有候选人的政党历史,以及政治选举影响(New York political studies(C)project[59])。

不仅关注本土地区历史,CCOH 还关注其他地区口述史研究,如中东、中欧地区。中东地区以突尼斯和伊朗为代表,如突尼斯过渡期口述历史收藏项目(Tunisian Transition Oral History Collection[60])记载了突尼斯革命(2010 – 2011 年)和过渡政府时期(2011 – 2014 年)历史,特别突出了总理梅赫迪·乔马(Mehdi Jomaa)任期突尼斯地区政治、教育、安全、经济发展。伊美关系项目(Iranian-American Relations Project[61])追溯了第二次世界大战后伊美外交关系发展,呈现了伊朗国王穆罕默德·雷扎·帕拉维(Mohammad Reza Pahlavi)任期间(1941 – 1979 年)伊朗地区历史及政治结构。中欧地区聚焦于匈牙利、阿根廷及奥地利。匈牙利项目[62]记载了匈牙利地区艺术、政治和国际关系的发展回忆;阿根廷项目[63]记录了阿根廷地区经济、工业、制造业相关技术变革;奥地利项目[64]记录了奥地利地区在音乐、文学、绘画和写作方面的发展。

换言之,以地区为主题的口述历史项目有计划、有步骤地开展,不仅扩大了地区历史研究视角,而且在一定程度上提升了地方口述历史研究价值。CCOH 在挽救和保护地方口述历史资源上做出了重要贡献,以口述历史方式记录地区政治、经济、文化、教育特色,以有效的本区发展布局联动他区机制协同,充分发挥了统筹规划作用。

5.4 以特定时期,具体运动/事件为主题——创新范式,观照家国时代主题

CCOH 以特定时期,具体运动/事件为主题的项目通常采集固定发生在某一年份或时间段内具有标志意义或舆论事件,针对社会热点或舆论度较高的问题进行调查、访谈,纵深挖掘该事件的起因、经过和结果,且多以战争背景为导向。

为纪念美国历史上唯一一次内战(也称为南北战争,1861 年 4 月 12 – 1865 年 4 月 9 日),内战百年纪念项目(Civil War Centennial Project)[65]记载了战争中有关奴隶制度废除和农民土地安置问题,该项目不仅维护了国家统一,为美国资本主义发展扫清了道路,同时也为美国跻身世界强国之列奠定了基础。飞虎队计划(Flying Tigers Project[66])是中美两国在第二次世界大战中并肩作战联合反抗日本法西斯侵略的精彩篇章,在中美关系史上占据重要位置,对增进两国信任、传承中美友好合作精神具有重要的启迪和指导作用。

除了纪念特定时期运动、发展事件,关注专题历史事件的史料收集,锁定重大历史事件是 CCOH 的另一独特视角。作为空前绝后的历史纪念事件,2001 年 9 月 11 日口述历史叙述与记忆项目(September 11, 2001 Oral History Narrative and Memory Project[67])由 5 个子项目组成,且访谈对象多元化,幸存者(September 11, 2001 Telling Lives Oral History Project[68])涵盖逃生的平民、救火队员、救护人员、街边的售卖者、死难家属、心理医生等,最大限度还原了美国东部时间 2001 年 9 月 11 日恐怖分子劫持 5 架民航客机撞击美国纽约世界贸易中心和华盛顿五角大楼的场景(美国纽约地标性建筑世界贸易中心双塔在内的 6 座建筑被完全摧毁)。为记载该事件的响应与恢复工作(September 11, 2001 Response and Recovery Oral History Project[69]),先后设立国家纪念馆、国家纪念广场和国家博物馆。同时自 2002 年设立"9.11 事件"纪念日以来,截至 10 周年,该项目已累计超过 900 小时的录音资料(其中包括 23 个小时的视频资料),访谈对象超过 600 人。迄今为止,351 人的 687 个小时的资料已对外开放,民众可通过调阅档案了解情况。

综上所述，鉴于特定时期，具体运动/事件本身的独特性，使得人们仅凭借文献和实物史料很难解读真实历史，在这种情况下，创新范式，借助相关亲历者和见证者的个人记忆和口头讲述就显得尤为重要。虽然口述历史本身具有流动性和不确定性，使得人们对它内容的真实性和客观性产生怀疑，但因为特定时期，具体运动/事件见证者的身份背景，也因为见证者个人记忆和历史事件之间的相互融合和转化，注定了见证者的口述历史也可以和文献资料一样，可以提供真实的和相对客观的史实事件还原。

6 结论

从上述 CCOH 项目主题解析来看，美国口述历史项目主题的丰富性得益于美国文化的厚重积淀。CCOH 项目建设已具备较为成熟的体系，大体分为 4 类，或立足人物，或以机构为视角，或聚焦地区，或以特定时期、具体运动/事件为主线展开，突破了以往偏重政治与上层阶级、"自上而下"的一贯式传统叙述范式。令人称道的是，除部分包含中国、中东和中欧历史外，多年来 CCOH 始终立足本土，以抢救地方历史、文化民俗作为首要目标，各行各业积极开展行动，全方位多领域立体渗透，激活历史人物文化基因，展现时代发展脉动，并且研究各有侧重。具体来说："聚人"——人物饱满性。人物是历史的创造者，通过各行各业人物群体的口述访谈映射了"以人为本"的人文思想；"聚构"——机构传承性。机构作为传承文化的重要载体，能唤醒与重塑社会记忆，多维构筑社会记忆体系。"聚地"——地区记忆性。口述历史作为能够较为完整地记录故事、历史与记忆的工具，其在记载地区历史、强化地区记忆方面具有广阔应用空间。"聚事"——事件省思性。口述历史并非直接"进入"日常生活的历史，而是重塑与追忆历史演进关键节点的重要历史时期与事件，通过事件记载唤醒事件经历。正是基于完整而丰富的主题体系，使得 CCOH 在完善自身研究成果的同时，培养出一批批口述历史专业人士，创造出一系列有价值的口述历史项目。

虽然从启动时间看，我国实践起点较美国晚了近 30 年，尚处于起步成长阶段。但作为中华文化的守护者和传承者，我国拥有相当丰富且数量庞大的人文资源亟待挖掘。可以预见，未来我国在口述历史项目体系构建、主题丰富度与细粒度、成果转化等方面大有可为。可从以下几个方面进行优化：

（1）组建专业团队，优化体系构建。与其他文献资源相比，口述历史资源内容广泛，载体形态多样，具有丰富馆藏和保护人类文化遗产等价值，以独特的人文属性传承历史记忆，因而其在采集、整理、传播与利用等阶段所需专业性极强。CCOH 通过招聘专业知识背景的采访者策划和参与访谈全过程，并成立专家顾问委员会，负责口述历史项目筹备。

目前，我国口述历史实践尚未形成完备的理论与方法，缺乏系统工作规范和相关人员配置。因此，我国口述历史项目开发主体应加强人力资源整合，优化实践体系。从教育培训层面，通过专业培训、实地调研、专题讲座、会议讨论等一系列方式组织知识理论与技能培训，涵盖组织采访、技术处理、项目统筹等多个环节；从人员配置层面，注重提升采访人员的专业知识素养、专题研究能力，组建专业口述历史团队，为口述历史研究输送人才，以专业化手段积极开展并组织完善各项工作，确保项目顺利开展。

（2）激活文化基因，展现时代脉动。体系构建是项目实践根基，选题规划是口述历史服务关键。CCOH 以"人－事－地－机构"四维层面涵盖了美国人文、政治、经济、教育、军事、文化等领域，多样化解读了大众生活、机构变迁、地区发展与事件感知，丰富了项目研究主题。

目前，我国口述历史项目选题范围相对狭窄，多以地方特色为主，以抢救保护地区文化为目的，易造成资源空缺、内容单一、缺乏创新性以及重复构建等弊端，不利于口述历史建设纵深扩散发展。因此延展选题范围，全方位多角度开展口述历史项目，丰富主题细粒度至关重要。从区域分配维度，可以承接本地政府、企业、专业协会及其他组织的委托课题，也可以跨地区，甚至跨国界合作共赢；在内容辐射广度上，不仅要关注政治、经济、宗教、艺术、民生等基本内容，而且要担当好中华文化的承载者和传播者，结合时代热点、文化基因，做好主题分类及挖掘，找寻项目实践切入点，还要在丰富我国口述历史研究内容的基础上逐步加注他国口述历史研究，增添内容饱满度，促进国际研究"大融合"。

（3）多元主体协同，联动发展机制。CCOH 联合相关院系、研究所和图书馆以线上线下推广、衍生品出售、多媒体互动及定期举办特色活动等方式，融合新媒体技术和数字网络技术，彰显出美国口述历史服务的人文关怀和为用户提供优质口述历史服务的初衷。

目前，我国口述历史项目实践尚停留在采集、存储、出版传播的初级阶段，缺少深度开发和知识挖掘。

虽然有构建数据库的积极性,但缺乏数字人文技术方法支撑,仅提供简单的项目检索和网上展览服务,服务方式保守,缺少吸引力。因此,我国口述历史项目开发主体如政府、高校、LAM(Library, Archive, and Museum)及社会组织等应积极借鉴CCOH联动参与机制,多元主体协同,突破传统运营和发展模式,打破资源孤岛,拓宽成果展示与传播渠道。

简言之,我国口述历史项目实践一方面要继续深入构建中国特色口述历史项目实践体系,另一方面要创造利于口述历史持续发展的社会环境,建立口述历史资源开发利用的联动机制,丰富项目研究主题,同时借助实体展览、AR/VR等数字人文技术等,面向公众分享个人记忆,推动文化遗产价值传播,助力文化遗产可持续发展。

参考文献:

[1] 杨祥银. 美国现代口述史学研究[M]. 北京: 中国社会科学出版社,2016:1.
[2] 哥伦比亚大学创新理论与实证跨学科中心(Interdisciplinary Center for Innovative Theory and Empirics, INCITE)[EB/OL]. [2020-02-08]. https://www.incite.columbia.edu/ccohr/.
[3] 哥伦比亚大学口述历史中心口述历史项目[EB/OL]. [2019-12-08]. https://oralhistoryportal.library.columbia.edu/project.php.
[4] 杨祥银. 试论口述史学的功用和困难[J]. 史学理论研究,2000(3):37-46.
[5] 张宏涛. 美国哥伦比亚大学口述历史工作研究及启示[J]. 图书情报工作,2019,63(3):123-128.
[6] 哥伦比亚口述历史研究中心(CCOHR)[EB/OL]. [2019-12-08]. http://www.ccohr.incite.columbia.edu/.
[7] 哥伦比亚口述历史档案中心(CCOHA)[EB/OL]. [2019-12-08]. https://incite.columbia.edu/ccohr? rq=CCOHA.
[8] 哥伦比亚大学口述历史中心研究指南[EB/OL]. [2019-12-08]. https://library.columbia.edu/content/dam/libraryweb/locations/ohro/The%20Oral%20History%20Bibliography%20--%20A%20CCOH%20Publication.pdf.
[9] 陈水湘. 美国高校口述历史工作调查与启示[J]. 图书馆建设,2018(5):34-41.
[10] 王玉龙. 基于案例分析的美国口述历史档案资源网络开发与利用[J]. 档案与建设,2017(2):21-24.
[11] 刘佳敏. 美国口述史特色及其借鉴[J]. 兰台世界,2017(1):28-30.
[12] 钟源,吴振寰. 美国10校图书馆口述历史工作调查分析[J]. 图书馆杂志,2019,38(6):83-91.
[13] 王鹏,范智新. 美国口述历史工作的特点及启示[J]. 中国档案,2019(6):76-77.
[14] 张钰. 口述历史项目中的法律伦理问题研究[D]. 南京:南京大学,2018.
[15] 季双琪. 美国图书馆口述历史资源库建设特色及启示[J]. 图书馆工作与研究,2018(11):40-45.
[16] 杨祥银. 数字化革命与美国口述史学[J]. 社会科学战线,2016(3):106-120.
[17] 陈方锐. 美国图书馆口述历史服务特点及启示[J]. 图书馆工作与研究,2018,1(4):5-10.
[18] 韦桥明. 美国科技领域口述档案项目资金来源研究[J]. 兰台世界,2016(10):32-35.
[19] BOYD D. Stories from the collection: Columbia University Oral History Research Office[J]. Oral history review,2001,28(2):137-138.
[20] LEVY J A. Apollo Theater Oral History Project and the Apollo Theater Oral History Project at C. S. 154 by Harlem Apollo Theater (review)[J]. Oral history review, 2014,41(1):134-136.
[21] VAUGHAN A G. African American cultural history and reflections on jung in the African diaspora[J]. Journal of analytical psychology,2019,64(3):320-348.
[22] DAVIS M P. Mexican voices/American dreams: an oral history of Mexican immigration to the United States[M]. New York:Henry Holt & Co., 1990.
[23] ABSHER A. The black musician and the white city: race and music in Chicago, 1900-1967[J]. Journal of the Illinois State Historical Society, 2015, 109(5):280.
[24] Lindsey Robinette Whitley. 非裔美国人口述历史项目[EB/OL]. [2019-12-08]. https://jewlscholar.mtsu.edu/bitstream/handle/mtsu/4893/Whitley_mtsu_0170N_10543.pdf? sequence=1&isAllowed=y.
[25] MILLER K A. Reading and writing the post-9/11 cop: trauma, personal testimony, and Jess Walter's the zero[J]. Arizona quarterly a journal of American literature culture and theory,2014,70(1):29-52.
[26] 李跃鹏, 金翠, 及俊川. 基于word2vec的关键词提取算法[J]. 科研信息化技术与应用, 2015,6(4):54-59.
[27] 曾志雄. 一种有效的基于划分和层次的混合聚类算法[J]. 计算机应用,2007,27(7):1692-1695.
[28] OLIVEIRA G, COUTINHO F, CAMPELLO R, et al. Improving k-means through distributed scalable metaheuristics[J]. Neurocomputing,2017,246:45-57.
[29] TAO H, LI J, LUO T, et al. Research on topics trends based on weighted K-means[C]//2017 7th IEEE international conference on electronics information and emergency communication. Macau: IEEE,2017:457-460.
[30] 沈思,梁晓静. 基于word2vec与k-means算法食品安全事件自动聚类研究[J]. 信息通信,2018,191(11):8-10.
[31] 哥伦比亚大学图书馆. 艾森豪威尔项目[EB/OL]. [2020-02-08]. https://clio.columbia.edu/catalog/4074583.
[32] 哥伦比亚口述历史研究中心. 巴拉克·奥巴马总统项目[EB/OL]. [2020-02-08]. https://www.ccohr.incite.columbia.edu/obama-presidency-oral-history-project.

[33] 哥伦比亚大学口述历史中心. 中国名人口述回忆录收藏项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/results. php? component_text = &projectTitle_t[ ] =%22Academia + Sinica + oral + history + collection. %22&projectTitle_tFacet[ ] =%22Academia + Sinica + oral + history + collection. %22.

[34] 哥伦比亚大学口述历史中心. 中国名人口述回忆录收藏项目－刘景山[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_13436782.

[35] 哥伦比亚大学口述历史中心. 中国名人口述回忆录收藏项目－贺国光[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_13435871.

[36] 哥伦比亚大学口述历史中心. 中国名人口述回忆录收藏项目－梅乔林[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_13436871.

[37] 哥伦比亚大学图书馆. 航空项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4072718.

[38] 哥伦比亚大学图书馆. 采矿工程师项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4076593.

[39] 哥伦比亚大学图书馆. 建筑项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4072675.

[40] 哥伦比亚大学口述历史中心. 诺贝尔科学研究奖项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/results. php? component_text = &projectTitle_t[ ] =%22Nobel + Laureates + on + scientific + research + project. %22&projectTitle_tFacet[ ] =%22Nobel + Laureates + on + scientific + research + project. %22.

[41] 哥伦比亚大学图书馆. 托马斯·阿尔瓦·爱迪生项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4074211.

[42] 哥伦比亚大学口述历史中心. 美国手工艺人项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_4075191.

[43] 哥伦比亚大学图书馆. 好莱坞电影产业项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4072703.

[44] 哥伦比亚大学口述历史中心. 大西洋慈善项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/results. php? component_text = &projectTitle_t[ ] =%22Atlantic + Philanthropies + Oral + History + Project. %22&projectTitle_tFacet[ ] =%22Atlantic + Philanthropies + Oral + History + Project. %22.

[45] 哥伦比亚大学口述历史中心. 黑人记者口述历史收藏项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_11561600.

[46] 哥伦比亚大学图书馆. 妇女历史和人口问题项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4074553.

[47] 哥伦比亚大学口述历史中心. 委内瑞拉劳工项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_13494501.

[48] 哥伦比亚大学图书馆. 纽约大学项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4076397.

[49] 哥伦比亚大学图书馆. 联邦百货商店项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4074565.

[50] 哥伦比亚大学图书馆. 世界银行项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4074552.

[51] 哥伦比亚大学图书馆. 美国物理教师协会项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4072421.

[52] 哥伦比亚大学图书馆. 哈里曼研究所项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/13273504.

[53] 哥伦比亚大学口述历史中心. 传教士计划[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_4074476.

[54] 哥伦比亚大学口述历史中心. 卡内基基金会项目－第一部分[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_4072539.

[55] 哥伦比亚大学图书馆. 卡内基基金会项目－第二部分[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/5459604.

[56] 哥伦比亚大学图书馆. 卡内基基金会项目－第三部分[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/10488425.

[57] 哥伦比亚大学图书馆. 纽约政治研究项目(A)[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4076400.

[58] 哥伦比亚大学图书馆. 纽约政治研究项目(B)[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4076398.

[59] 哥伦比亚大学图书馆. 纽约政治研究项目(C)[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4076399.

[60] 哥伦比亚大学图书馆. 突尼斯过渡期口述历史收藏项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/13275174.

[61] 哥伦比亚大学口述历史中心. 伊美关系项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/document. php? id = ldpd_4077015.

[62] 哥伦比亚大学图书馆. 匈牙利项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4076118.

[63] 哥伦比亚大学图书馆. 阿根廷项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4072586.

[64] 哥伦比亚大学图书馆. 奥地利项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4072697.

[65] 哥伦比亚大学图书馆. 内战百年纪念项目[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4074005.

[66] 哥伦比亚大学图书馆. 飞虎队计划[EB/OL]. [2020－02－08]. https://clio. columbia. edu/catalog/4074563.

[67] 哥伦比亚大学口述历史中心. 2001 年 9 月 11 日口述历史叙述与记忆项目[EB/OL]. [2020－02－08]. https://oralhistoryportal. library. columbia. edu/results. php? component_text = &projectTitle_t[ ] =%22September + 11%2C + 2001 + oral + history + narrative + and + memory + project. %22&projectTitle_tFacet

[ ] = %22September + 11%2C + 2001 + oral + history + narrative + and + memory + project. %22.

[68] 哥伦比亚大学口述历史中心. 2001 年 9 月 11 日叙述者口述历史项目[EB/OL]. [2020 - 02 - 08]. https://oralhistoryportal. library. columbia. edu/results. php? component_text = &projectTitle_t[ ] = %22September + 11%2C + 2001 + telling + lives + oral + history + project% 3A% 22&projectTitle _ tFacet [ ] = % 22September + 11%2C + 2001 + telling + lives + oral + hishttps://oralhistoryportal. library. columbia. edu/results. php? component_text = &projectTitle_t[ ] = %22September + 11%2C + 2001 + telling + lives + oral + history + project% 3A% 22&projectTitle_tFacet [ ] = %22September + 11%2C + 2001 + telling + lives + oral + history + project%3A%22 tory + project%3A%22.

[69] 哥伦比亚大学口述历史中心. 2001 年 9 月 11 日响应与恢复口述历史项目[EB/OL]. [2020 - 02 - 08]. https://oralhistoryportal. library. columbia. edu/results. php? component _ text = &projectTitle_t[ ] = %22September + 11%2C + 2001 + response + and + recovery + oral + history + project. % 22&projectTitle_tFacet [ ] = %22September + 11%2C + 2001 + response + and + recovery + oral + history + project. %22.

作者贡献说明：

王阮：负责数据分析、论文撰写与修改；
邓君：提出论文整体研究思路与框架；
孙绍丹：负责数据处理；
宋先智：负责数据收集；
宋雪雁：负责数据收集。

Cluster Analysis and Theme Deconstruction of the Project of Columbia Center for Oral History

Wang Ruan　Deng Jun　Sun Shaodan　Song Xianzhi　Song Xueyan
School of Management, Jilin University, Changchun 130022

Abstract: [**Purpose/significance**] To explore the characteristics of Columbia Center for Oral History (CCOH) project and deconstruct the theme of the project from a new perspective, so as to provide a reference for enriching the theme of oral history project in China. [**Method/process**] 9 586 oral history projects of CCOH were selected as the research objects, data were captured by Python, and 3 589 projects (5 997 projects without brief introduction) were filtered through Word2Vec and K-means, and the main theme of the projects was deconstructed. [**Result/conclusion**] The project of CCOH consists of four thematic levels: ①Taking people as the theme - all inclusive, radiating to all walks of life; ②taking institutions as the theme - learning from all strengths, giving full play to their own functional advantages; ③taking regions as the theme - overall planning, linkage of regional system construction; ④taking specific periods, specific movements / events as the theme - innovation paradigm, reflecting on the theme of home and country times.

Keywords: CCOH　project theme　Word2Vec　K-means